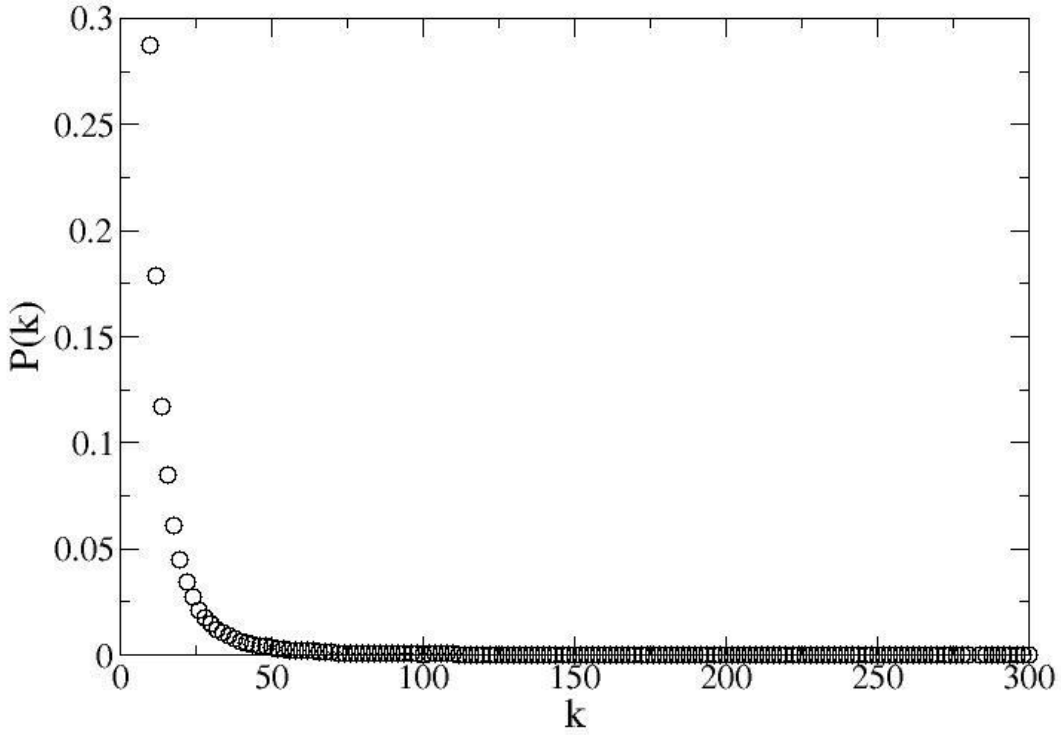
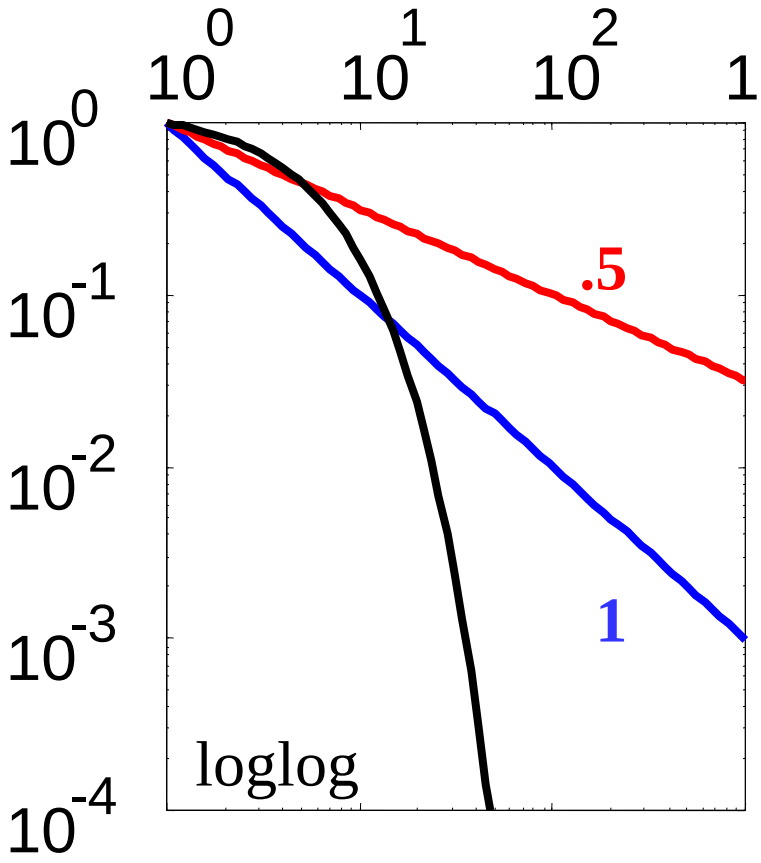
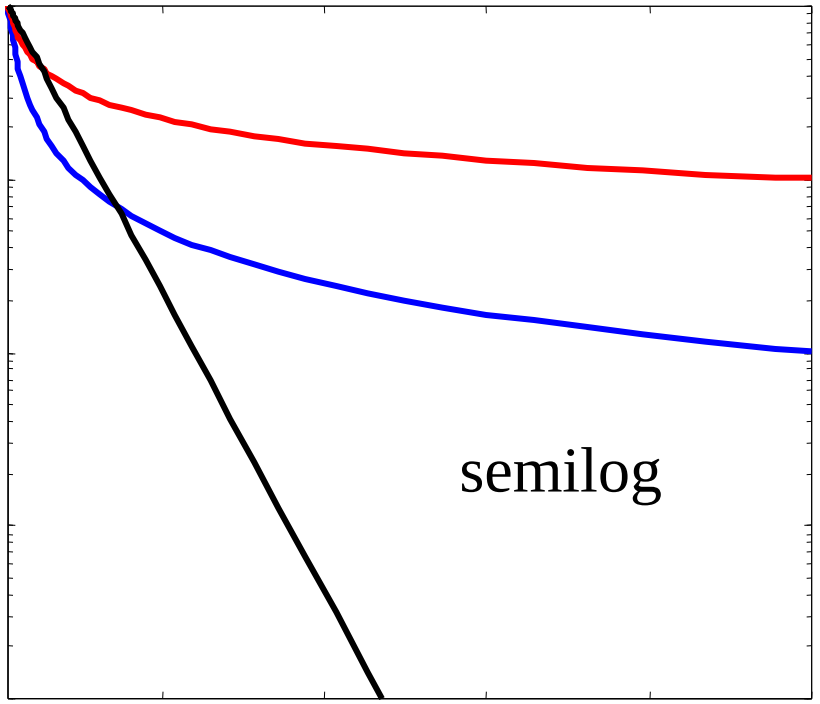
Properties of real networks: degree distribution



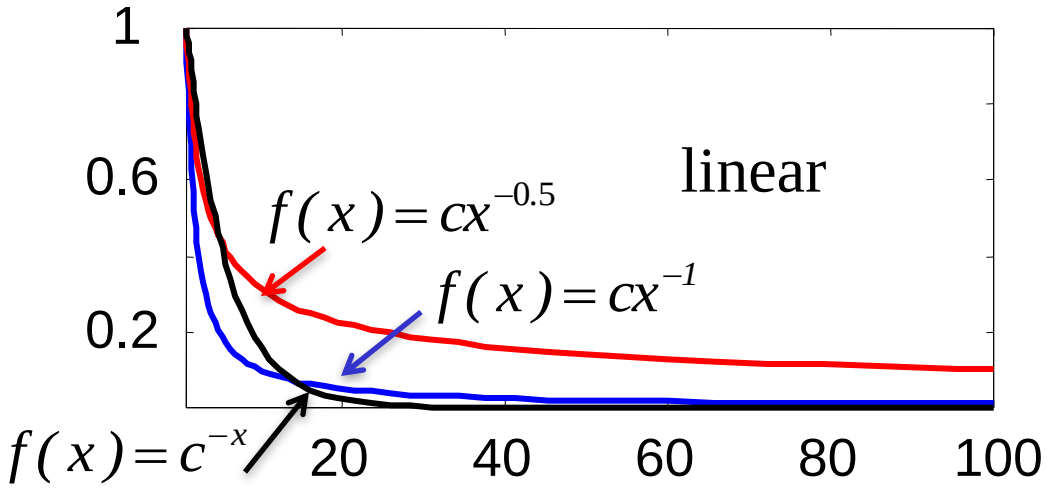
Nodes with small degrees are most frequent.
The fraction of highly connected nodes decreases, but is not zero.
Look closer: use a logarithmic plot.



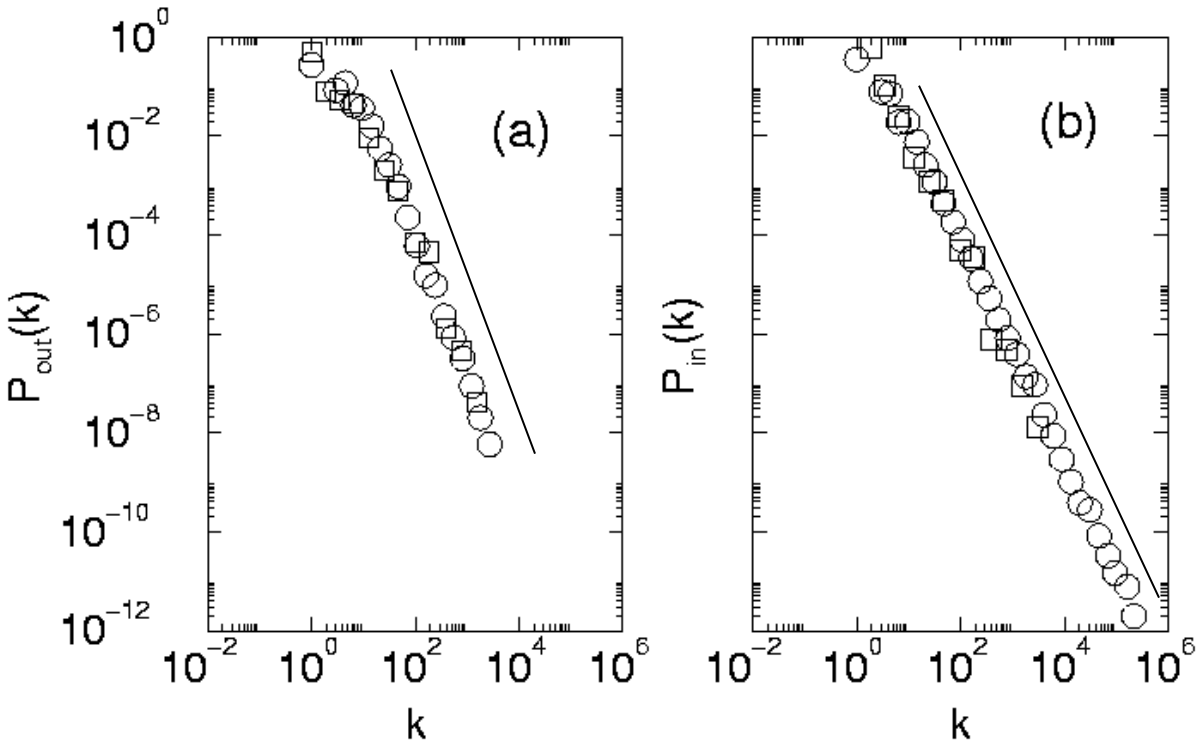


Plotting power laws and exponentials

Note: these are plots of functions and not degree distributions



In- and out-degree distribution of the WWW



nodes: webpages

edges: hyperlinks

$$P_{out}(k) \approx k^{-2.45}$$

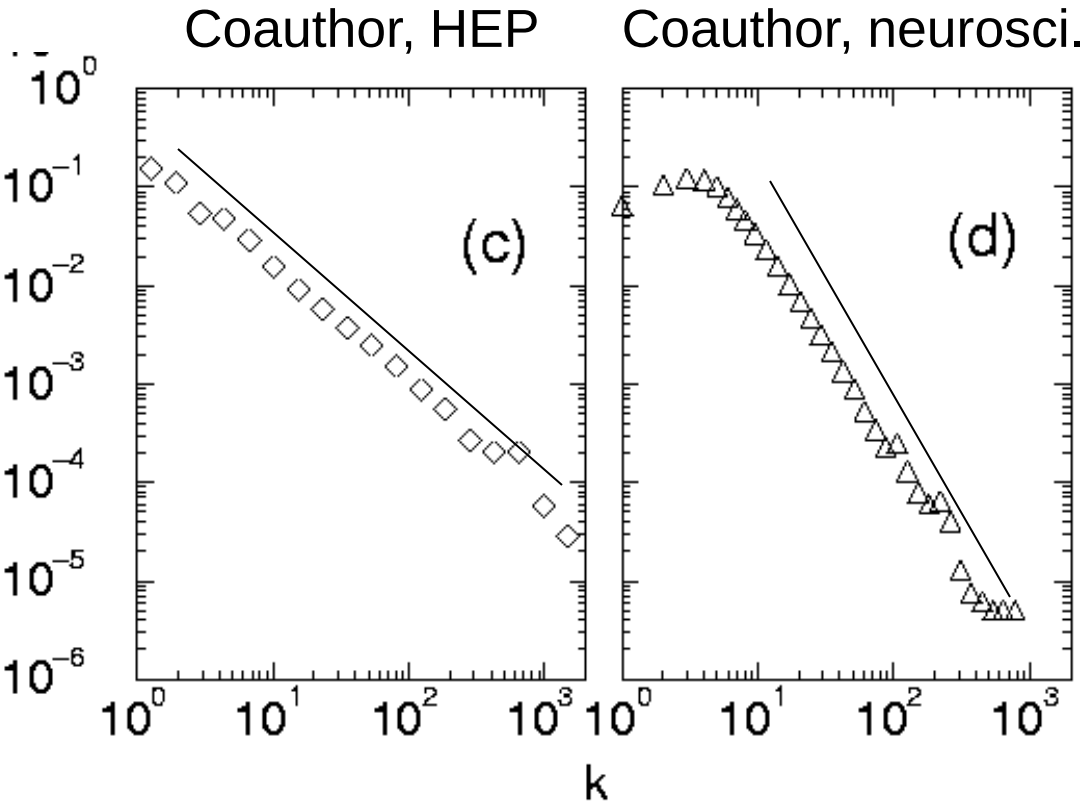
$$P_{in}(k) \approx k^{-2.1}$$

Usage: the degree distribution scales as a power law

R. Albert, H. Jeong, A.-L. Barabási, Nature 401, 130 (1999)

A. Broder *et al*., Comput. Netw. 33, 309 (1999)

Degree distributions in networks of science collaborations



$P(k) \approx k^{-1.2}$

$P(k) \approx k^{-2.1}$

M. E. J. Newman, Phys. Rev. E 64, 016131 (2001)

A.-L. Barabási et al., cond-mat/0104162 (2001)

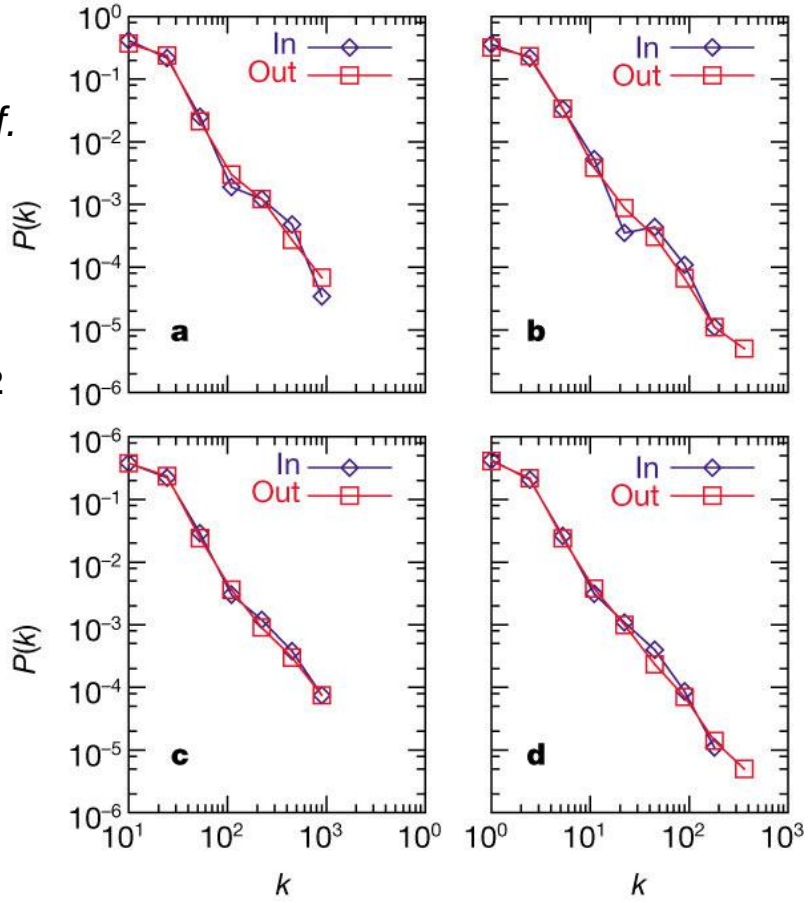
Metabolic networks have a power-law degree distribution

Archaeoglobus f.

$$P_{in}(k) \approx k^{-2.2}$$

$$P_{out}(k) \approx k^{-2.2}$$

C. elegans



E. coli

bipartite

nodes: metabolites,

reactions

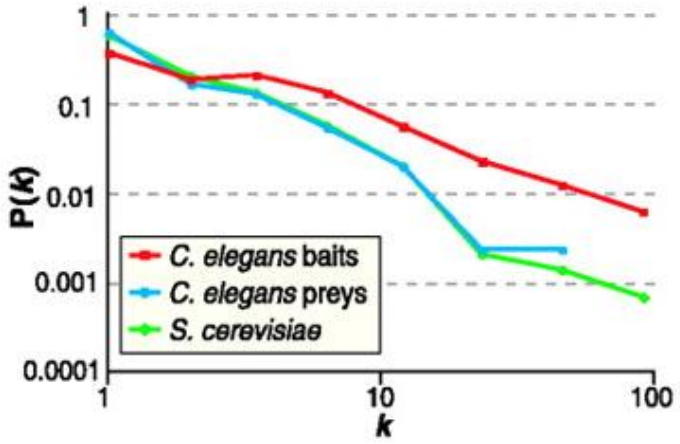
directed edges,

out: reactant (substrate)

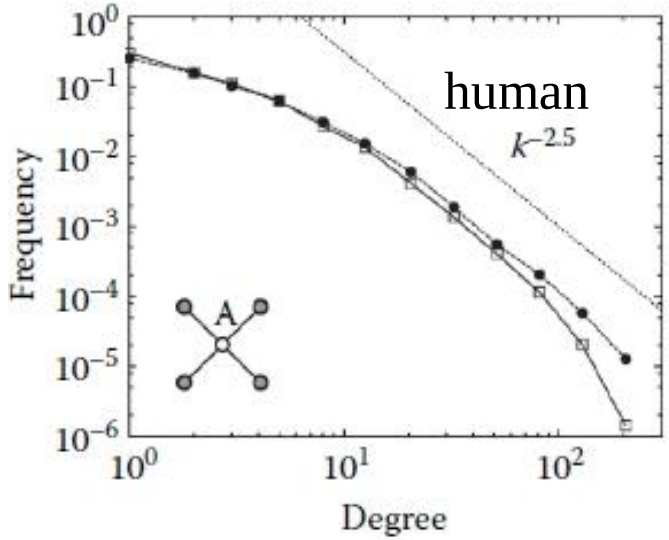
in: product of reaction

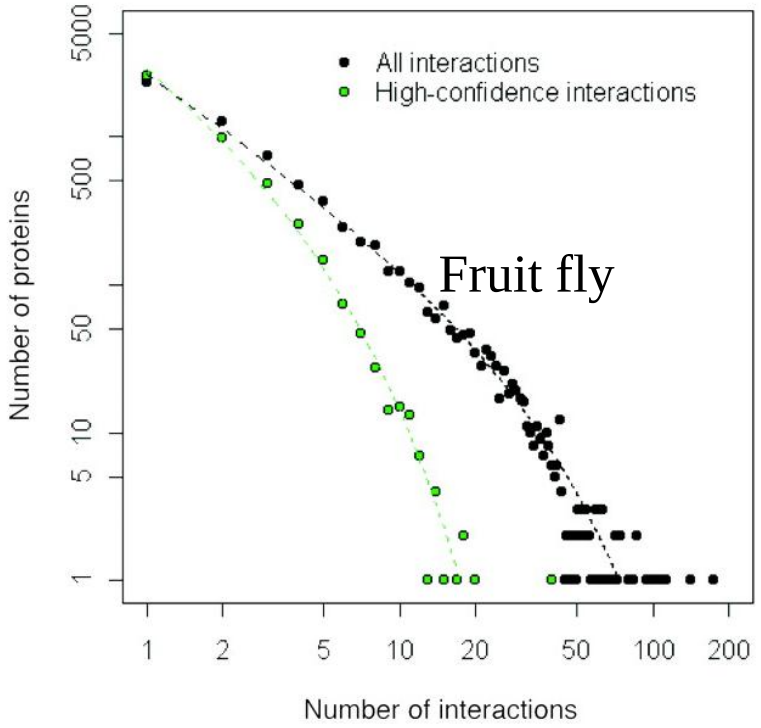
H. Jeong et al., Nature 407, 651 (2000)

Degree distribution of protein networks



$$P(k) \approx Ak^{-\gamma}$$

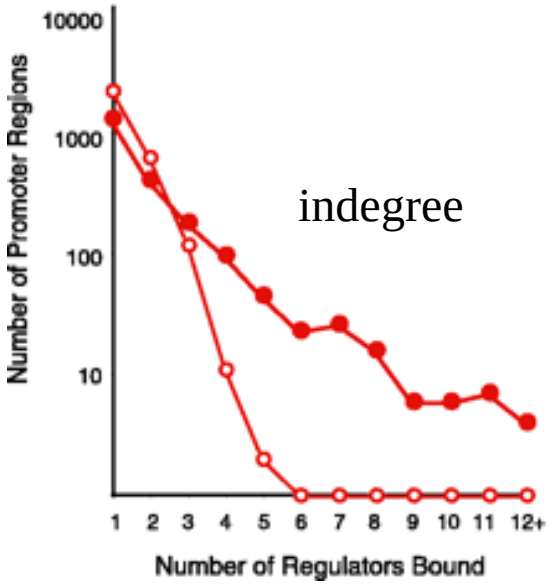




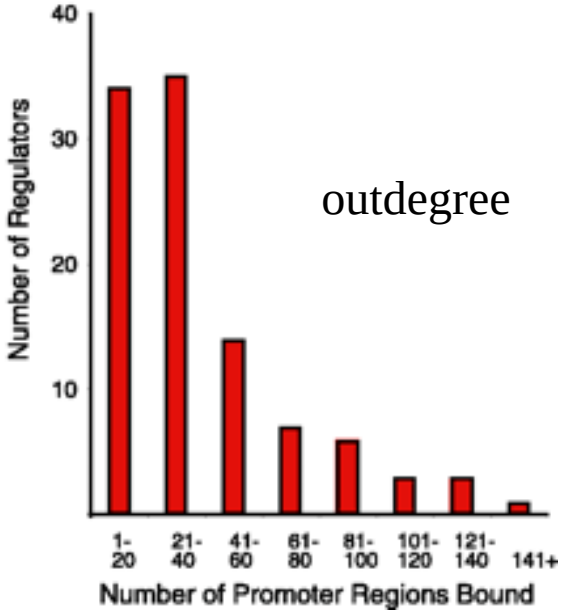
$$P(k) \approx Ak^{-\gamma} \exp(-\beta k)$$

Giot et al. Science 2003 – *Drosophila m.*
Li et al. Science 2004 – *C. elegans*
Rual et al. Nature 2005 – human
Stelzl et al. Cell 2005 - human

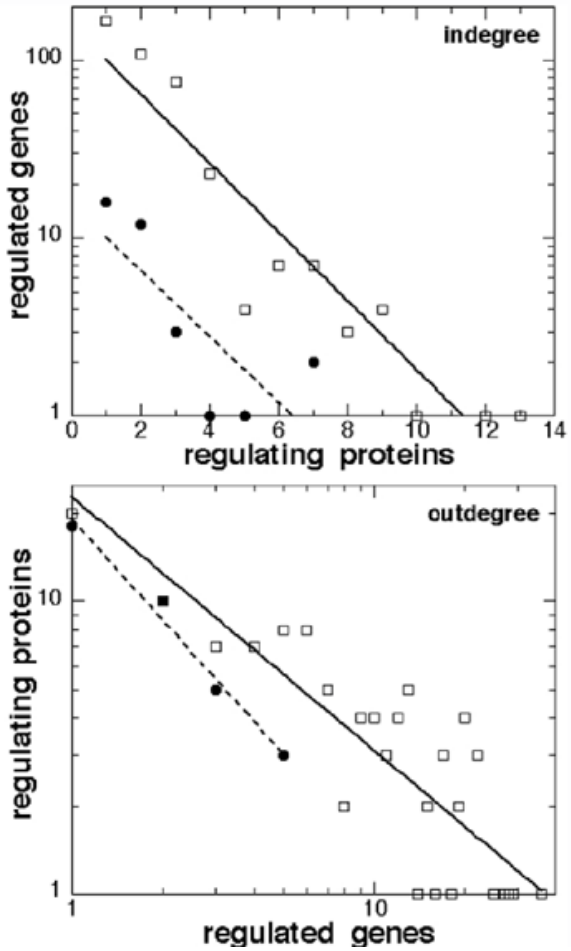
Gene regulatory networks' out-degree distribution long - tailed, in-degree distribution more limited



indegree



outdegree



Guelzim et al, Nature Genetics 31, 60 (2002)
Lee et al, Science 298, 799 (2002)

S. cerevisiae

Cleaning up degree distributions

Often it is difficult to determine the best fit to the points that make up a degree distribution.

Methods of data cleanup for decreasing degree distributions:

1. logarithmic binning: bin the k range; use bins of exponentially increasing size
2. Display the cumulative degree distribution

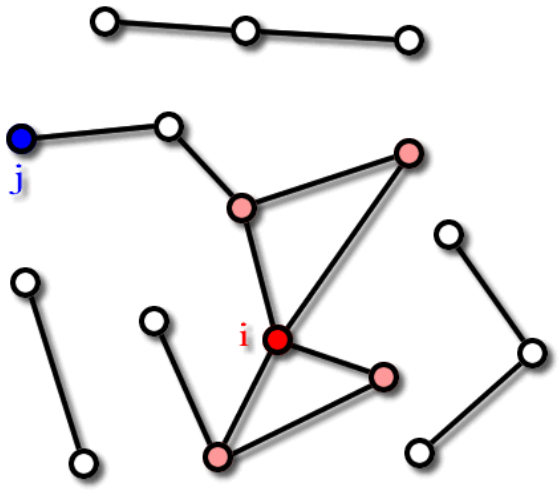
$$P(k > K) = 1 - P(k \leq K)$$

3. Construct a rank-degree plot wherein nodes are ranked in the decreasing order of degree.

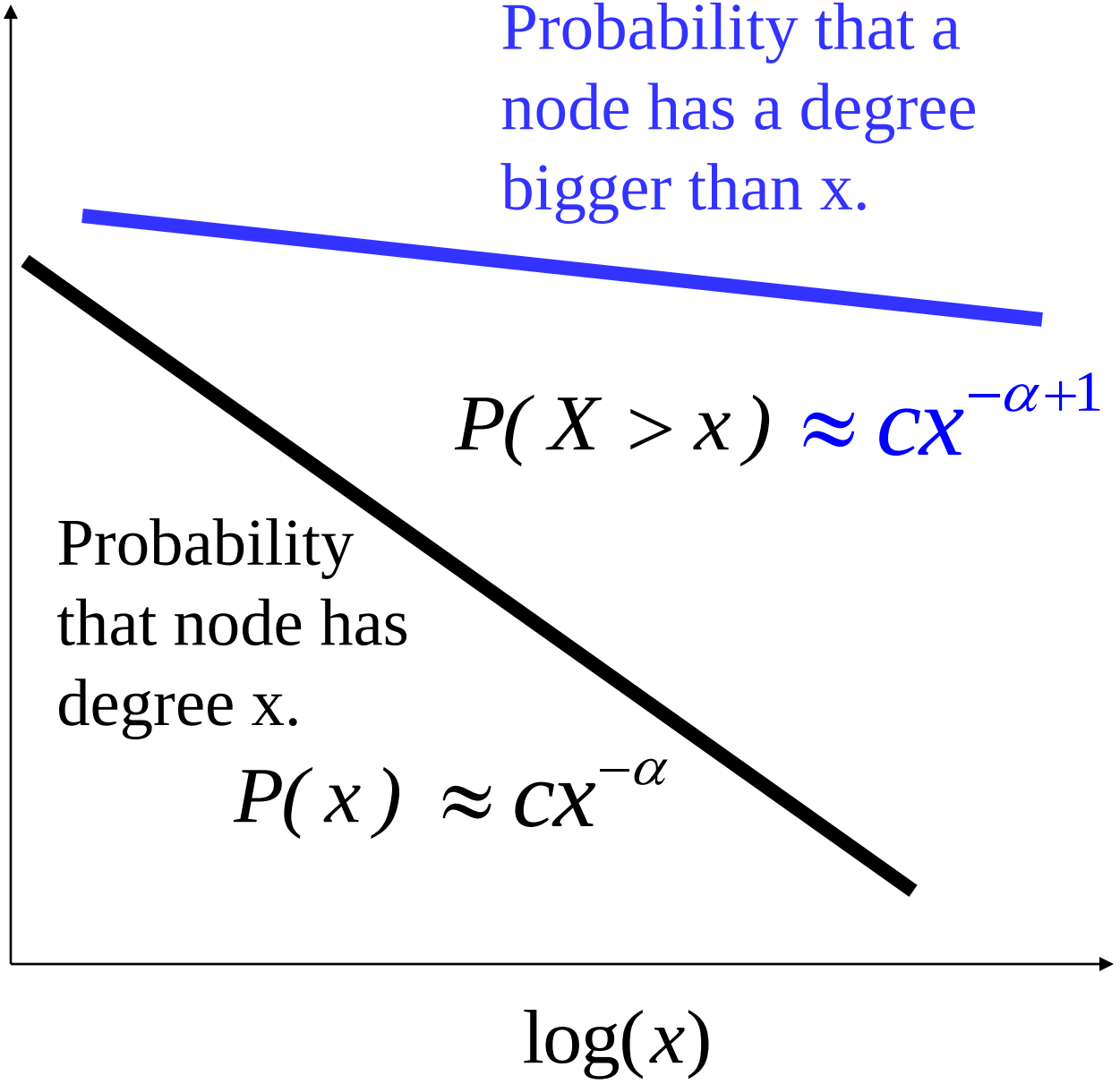
J. Wu et al., Comp. Bio. and Chem. 32, 1 (2008)

Ex. Determine the degree distribution and cumulative degree distribution of the graph on the right. Construct its rank-degree plot.

$$P(k > K) = 1 - P(k \leq K)$$

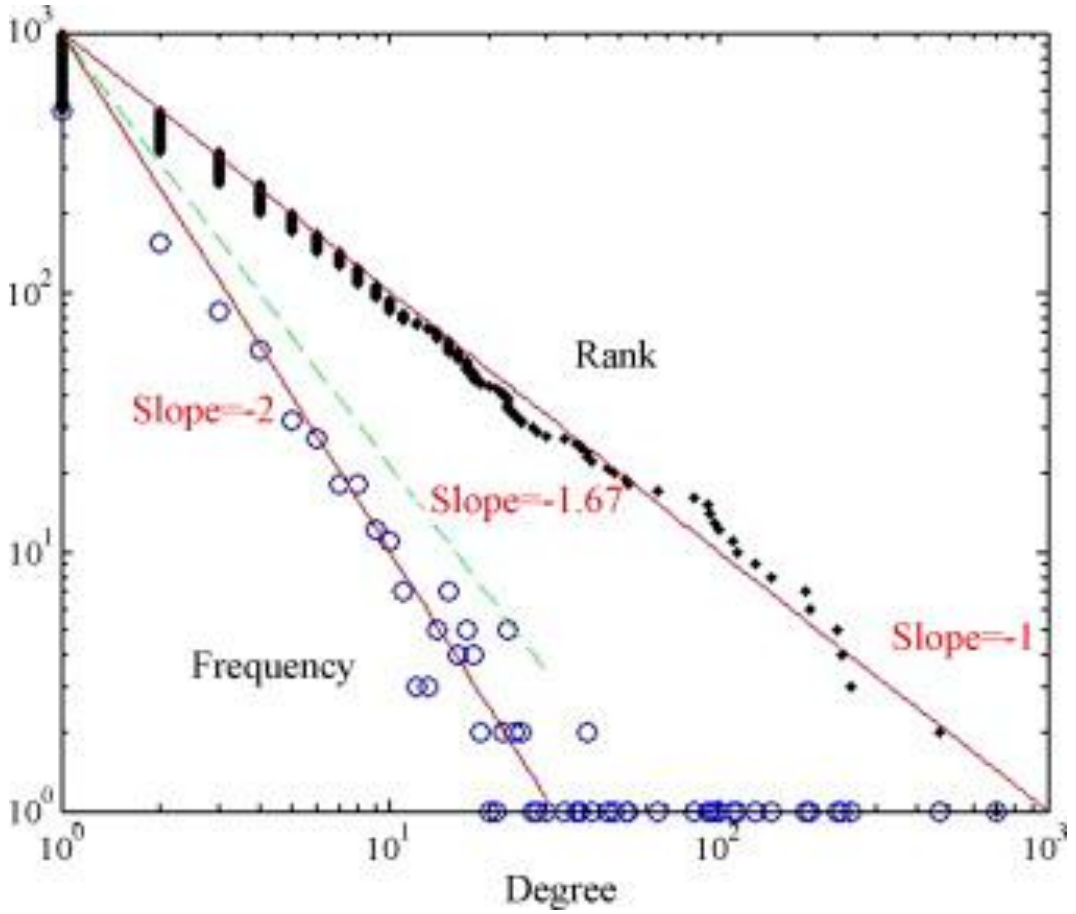


If the (noncumulative) degree distribution aligns with a power law with exponent α>1, the cumulative degree distribution will align with a power law with exponent α-1. Does not apply for α=1!



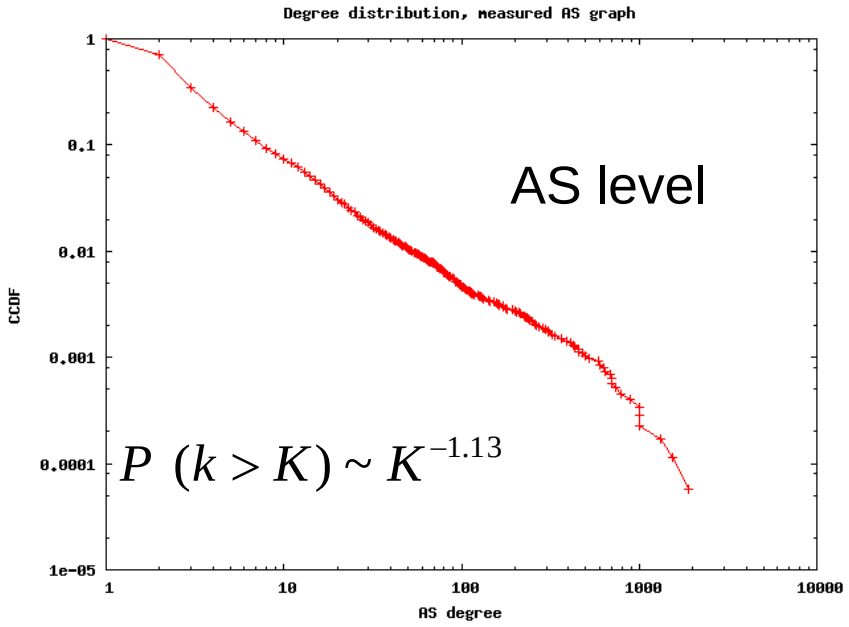
Degree distribution and rank-degree plot

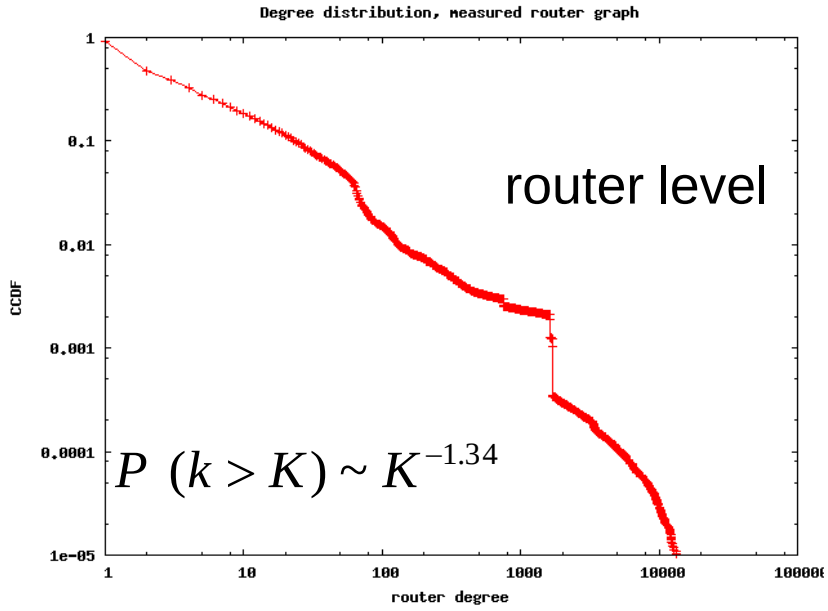
If the (noncumulative) degree distribution aligns with a power law with exponent α>2, the rank-degree distribution will align with a power law with exponent α-1.



J. Wu et al., Comp. Bio. and Chem. 32, 1 (2008)

Cumulative degree distributions of the Internet

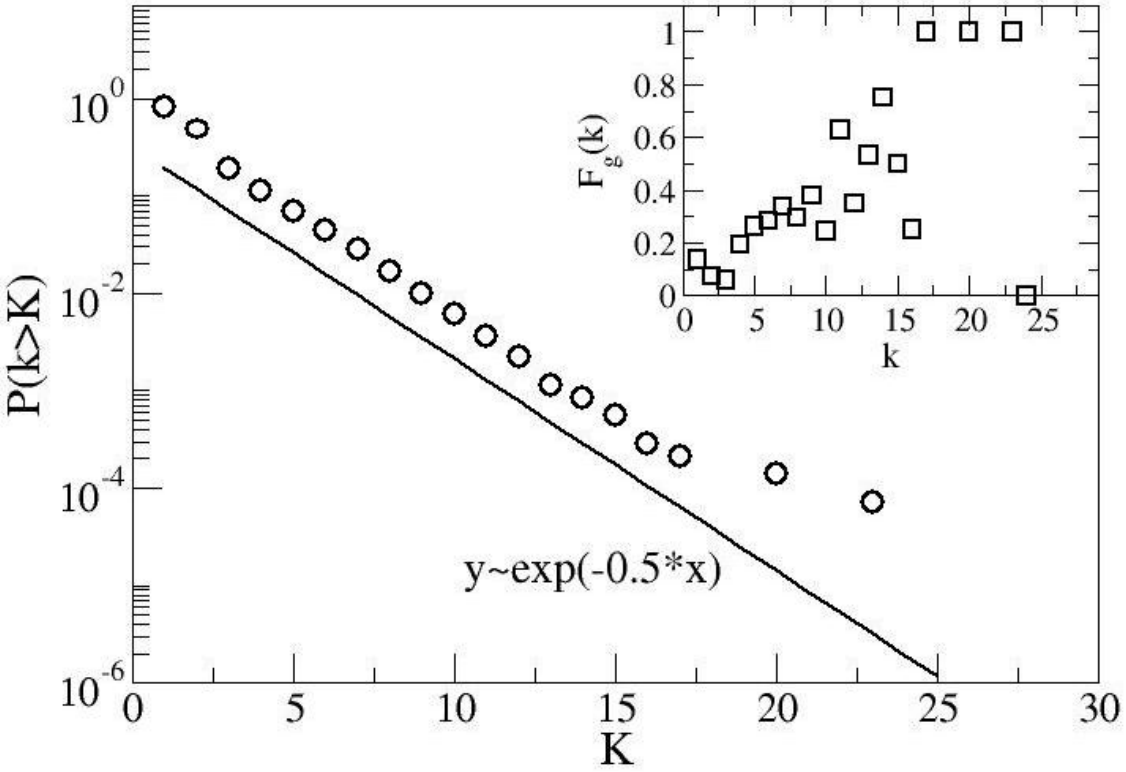




CCDF: complementary cumulative distribution function, P(k>K)

CAIDA, http://www.caida.org/research/topology/generator/

Power grid has exponential degree distribution



nodes: generators,

power stations

edges: power lines

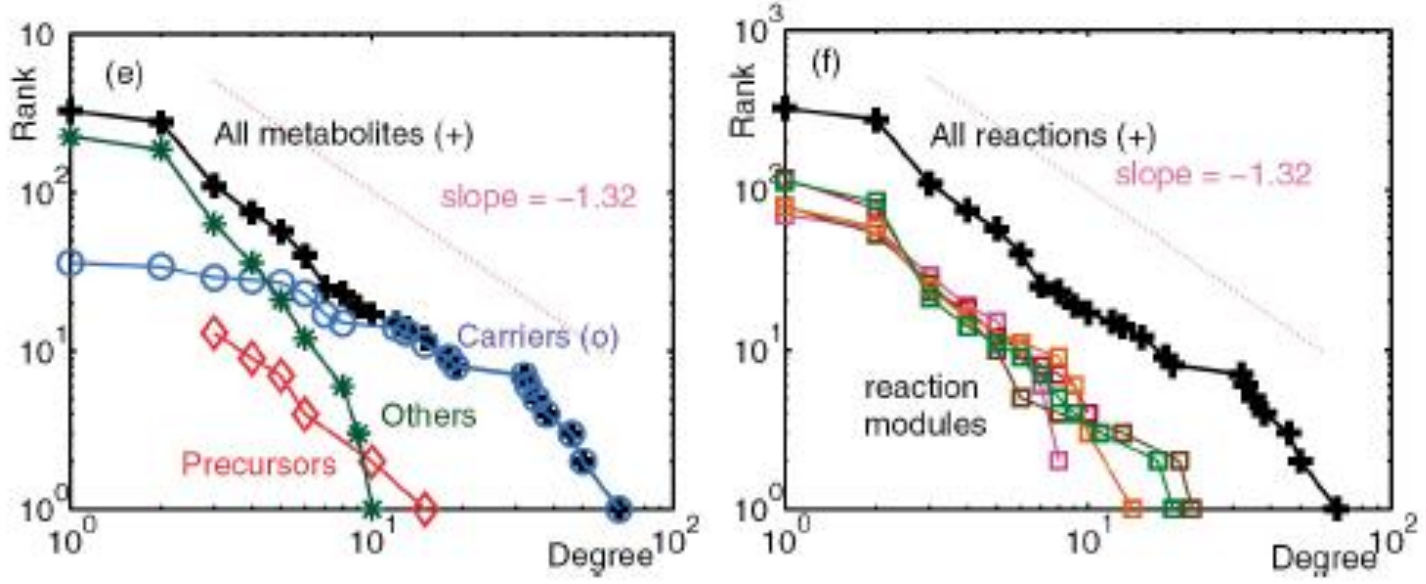
$$P\ (k > K) \propto exp(0.5K)$$

R. Albert, I. Albert, G. L. Nakarado, Phys. Rev. E 69, 025103(R) (2004)

Degree distributions in metabolite and reaction networks

Construct non-directed projections to metabolite and reaction networks

Rank vs. degree plot, similar to P(k>K).
The degree exponent γ=|slope|+1

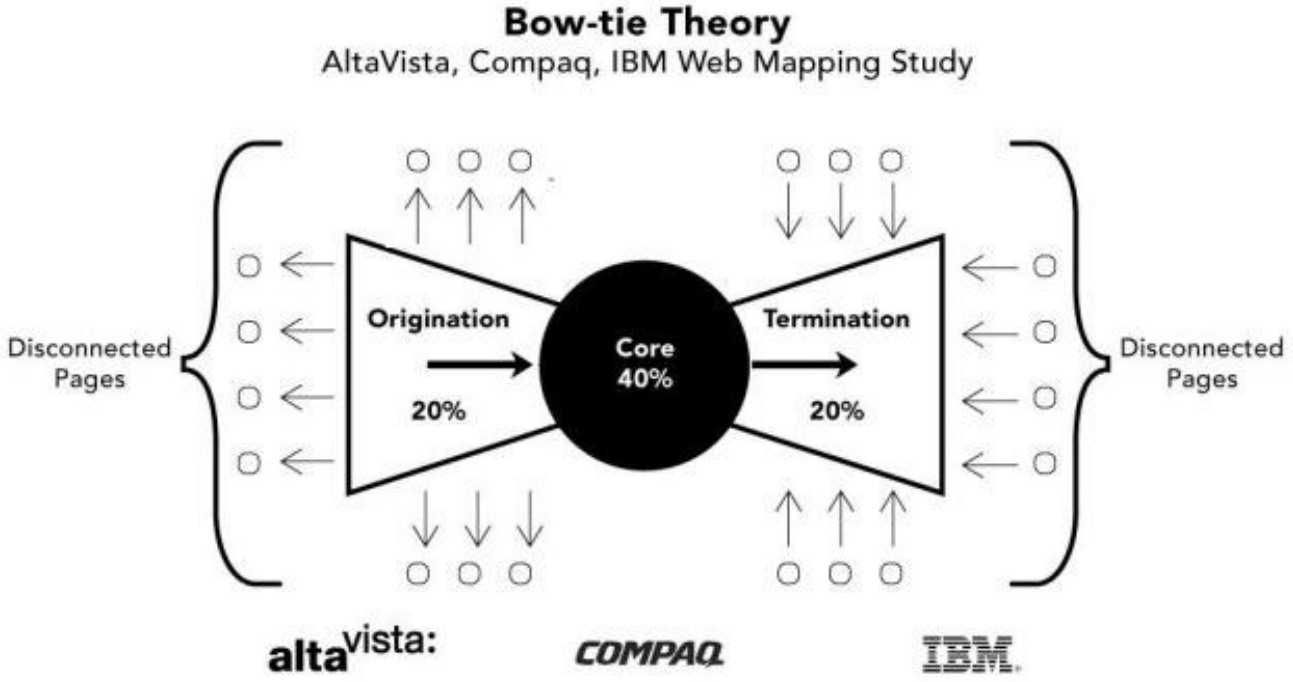


Undirected substrate network Undirected reaction network

Tanaka, Phys. Rev Lett. 94, 168101 (2005)

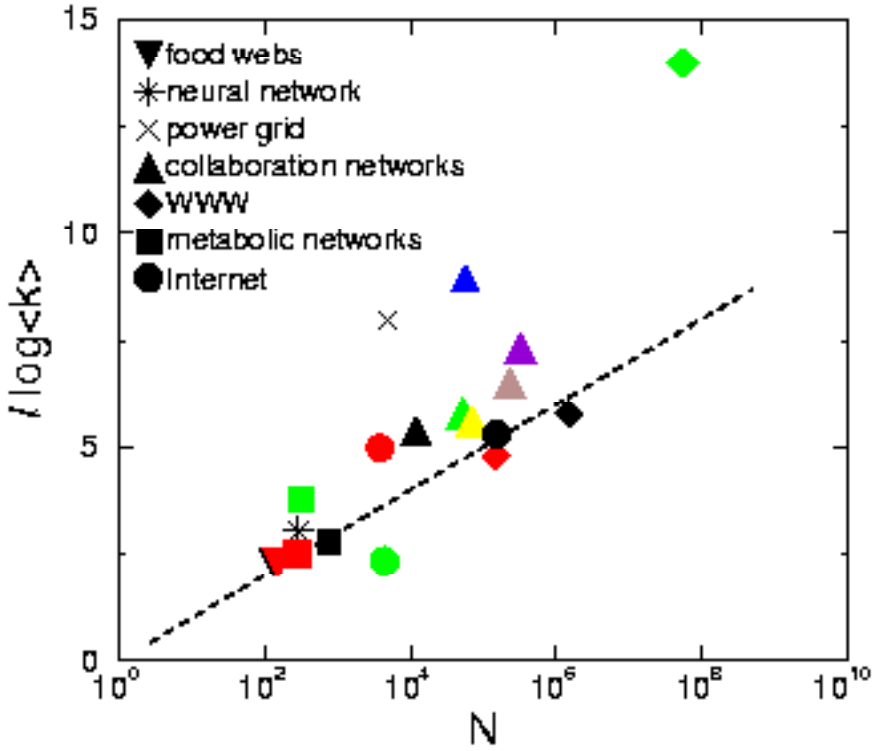
Bow-tie structure of the WWW

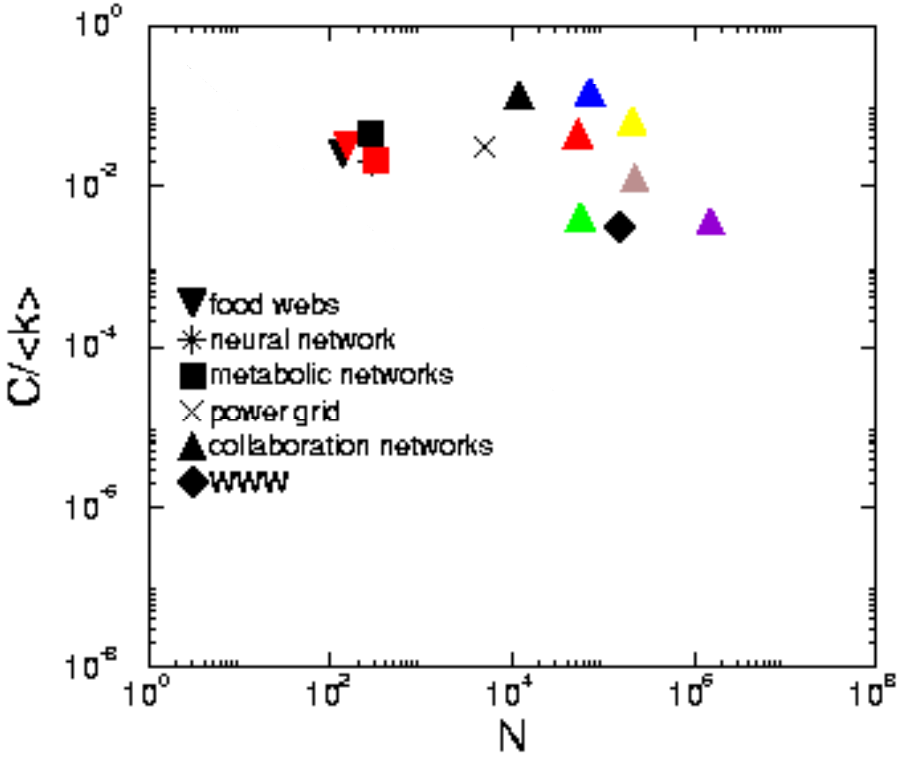
Network has >200 million webpages, >1.5 billion hyperlinks
Largest strongly connected component (Core) <40% of network
In-component (Origination) ~20%
Out-component (Termination) ~20%



Broder et al, Comput. Netw. 33, 309 (2000).

Average path length and average clustering coefficient in real networks



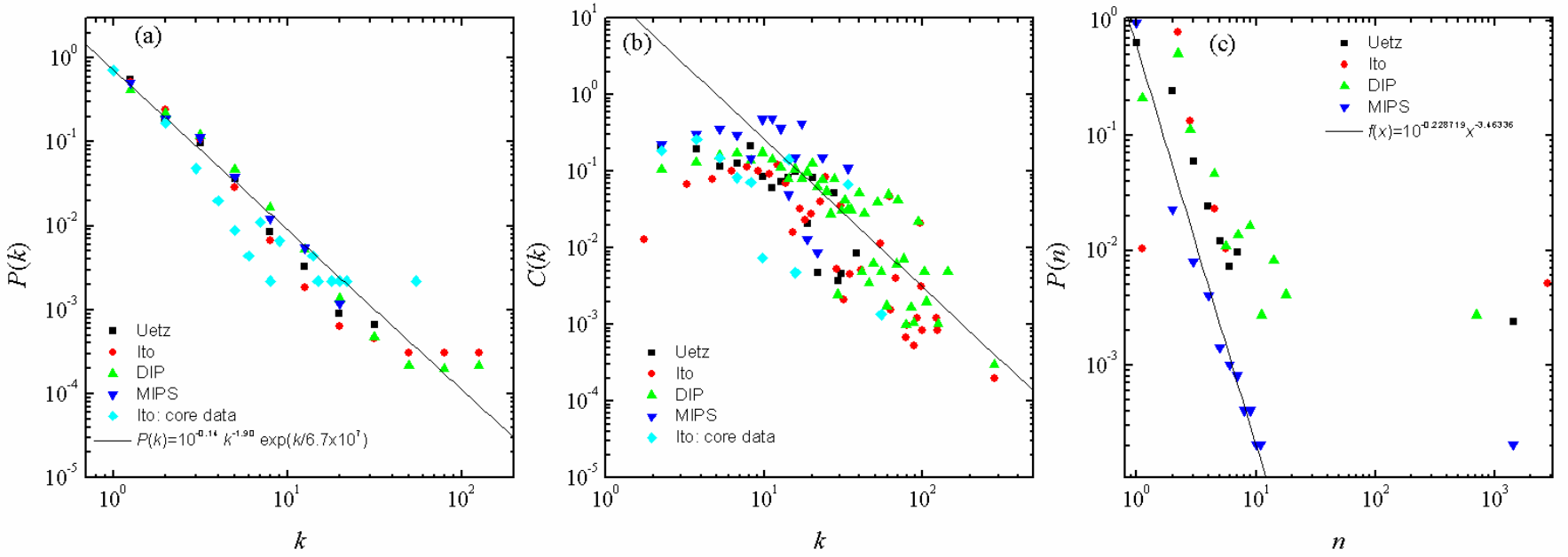


$$\langle l \rangle \approx \frac{log\ N}{log \langle k \rangle}$$

$$\langle C \rangle \propto \langle k \rangle$$

Apparent scaling with the network size and average degree - as though these different networks were members of the same family.

Comparison of yeast interaction networks



Degree distribution

$$P(k) \sim k^{-2.5}$$

Clustering coefficient

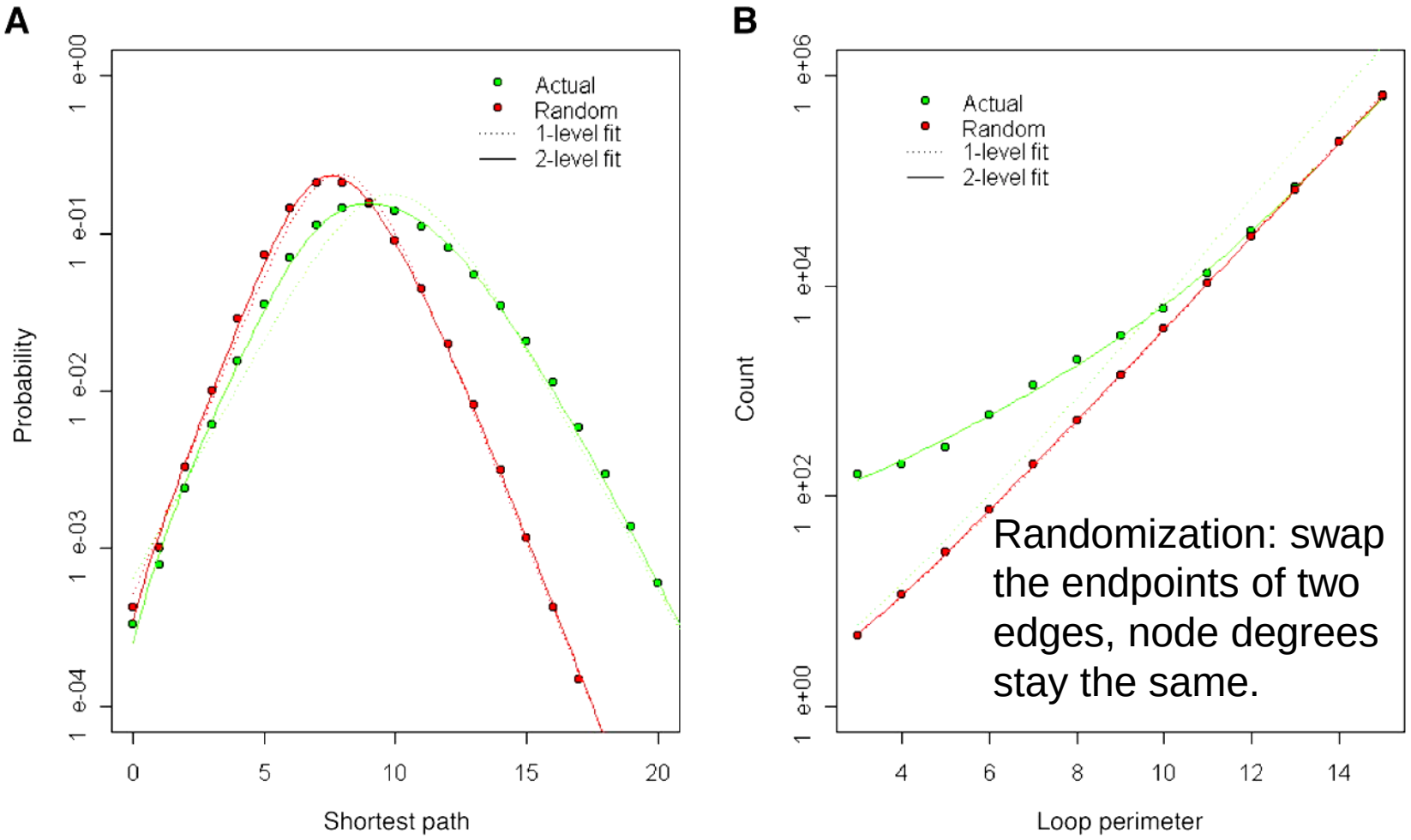
$$C(k) \sim k^{-2}$$

Connected components

$$p(n) \sim n^{-3.4}$$

Yook, Oltvai and Barabási, Proteomics 4, 928 (2004)

Paths in Drosophila protein interaction network



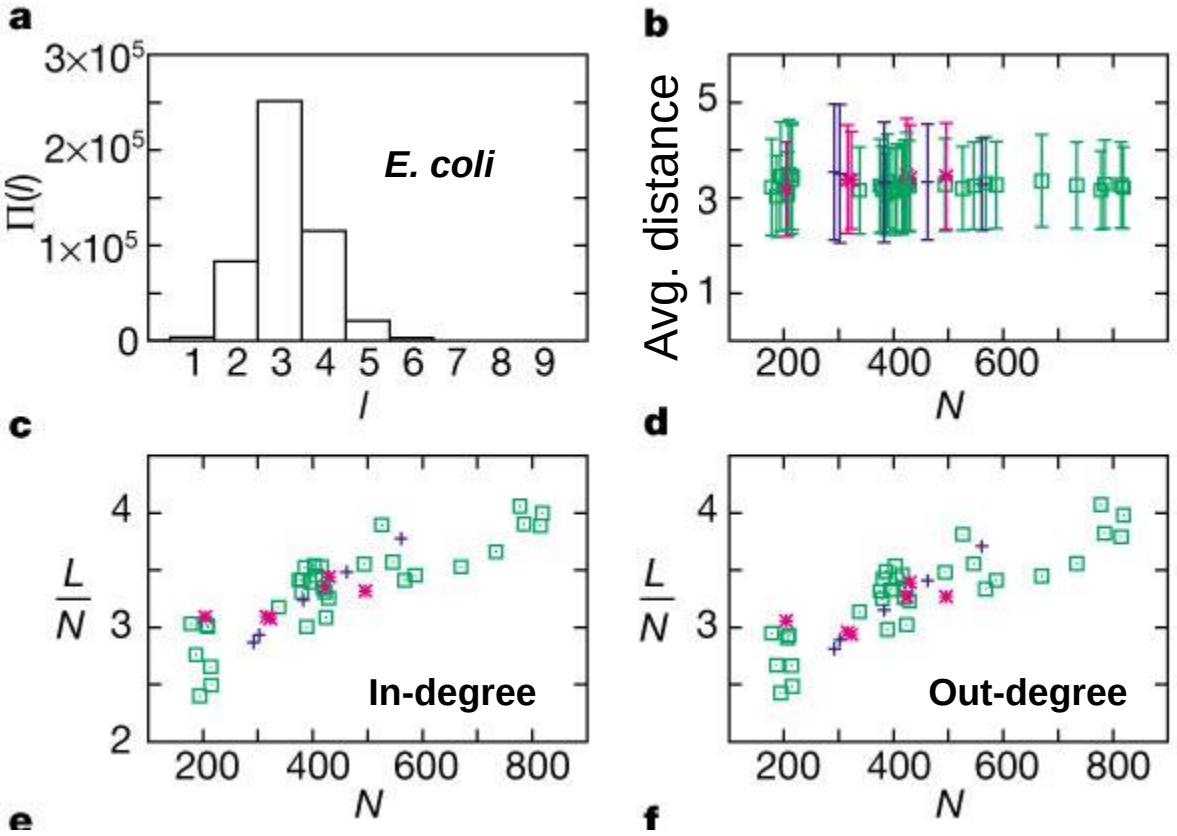
More long paths, but also more short cycles, than in randomized network.

Distances in Metabolic Networks

Paths defined to connect substrates (reactants) to products, the average is calculated on the reachable pairs only.

Distance distribution

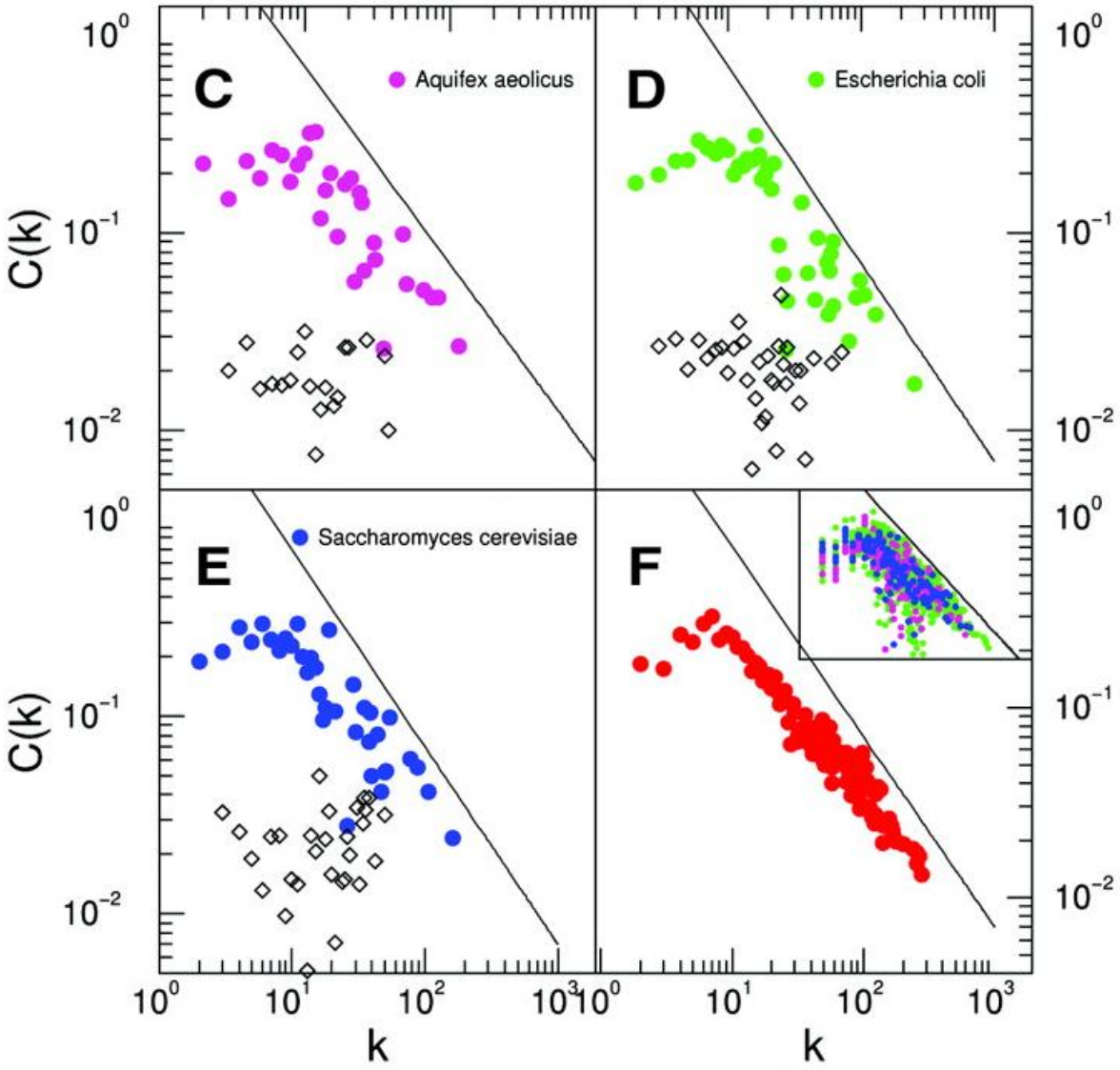
Average degree



Relatively small and constant average distance across organisms

H. Jeong *et al.*, Nature 407, 651 (2000)

Clustering-degree relation in metabolic networks



Average clustering coefficient of nodes with degree k

Open symbols: a model with the same degree distribution

Straight line: $C(k) \sim k^{-1}$

Ravasz et al., Science 297, 1551 (2002)

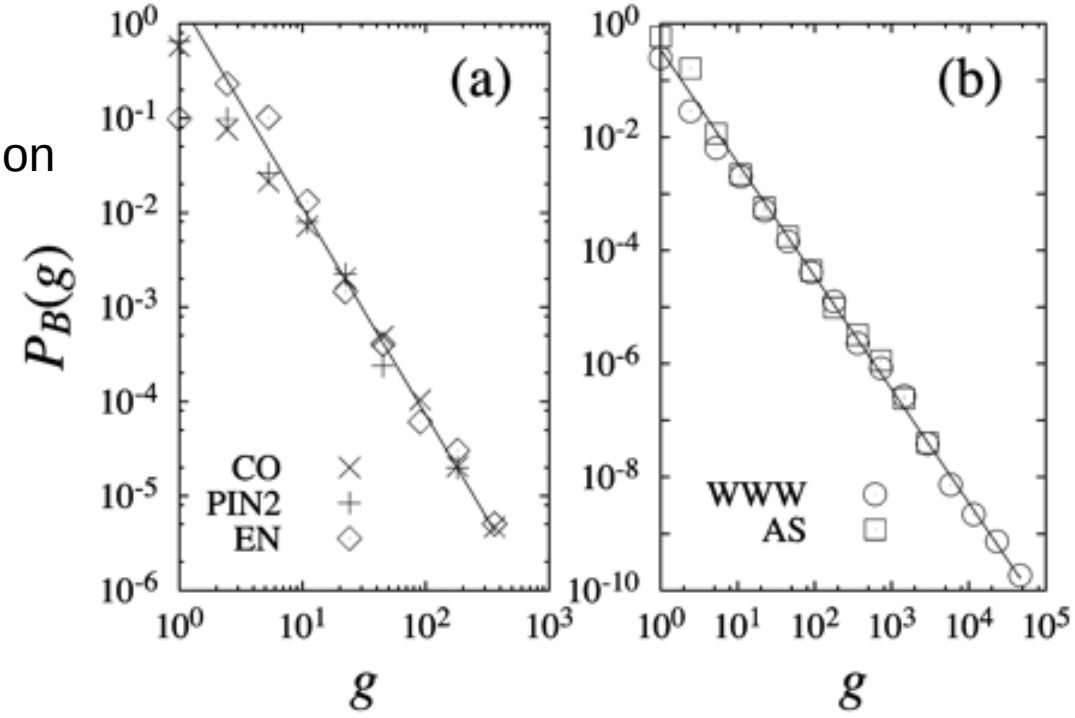
Distribution of betweenness centrality

Coauthorship

Protein interaction

Metabolic netw.

$$P_B(g) \approx g^{-2.2}$$



World-wide Web

Internet (AS level)

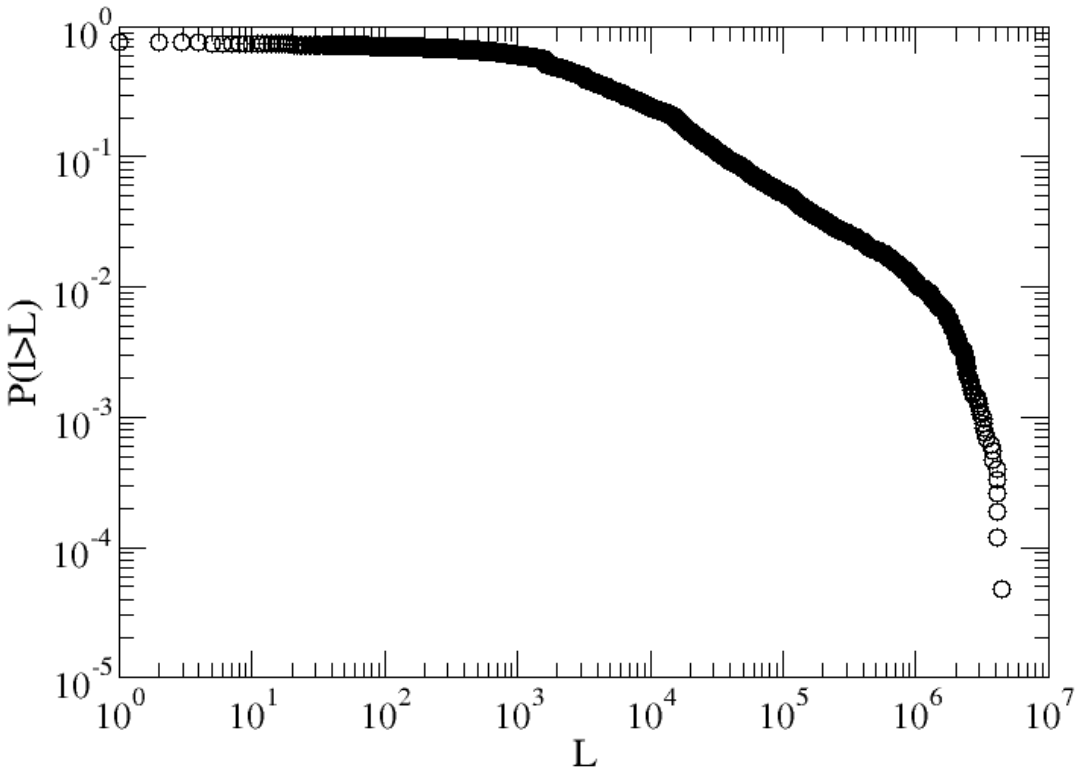
$$P_B(g) \approx g^{-2}$$

K. I. Goh et al., PNAS 99, 12583 (2002)

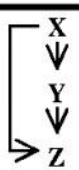
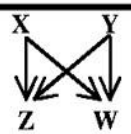
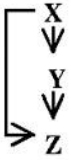
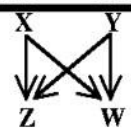
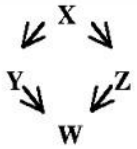
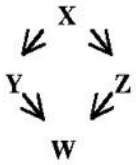
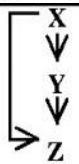
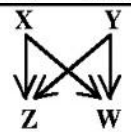
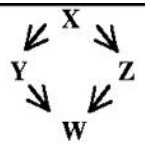
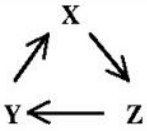
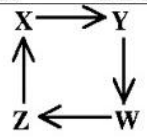
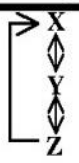
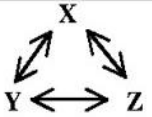
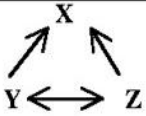
Betweenness centrality (load) distribution of the power grid

$$P(l > L) \approx (2500 + L)^{-0.7}$$

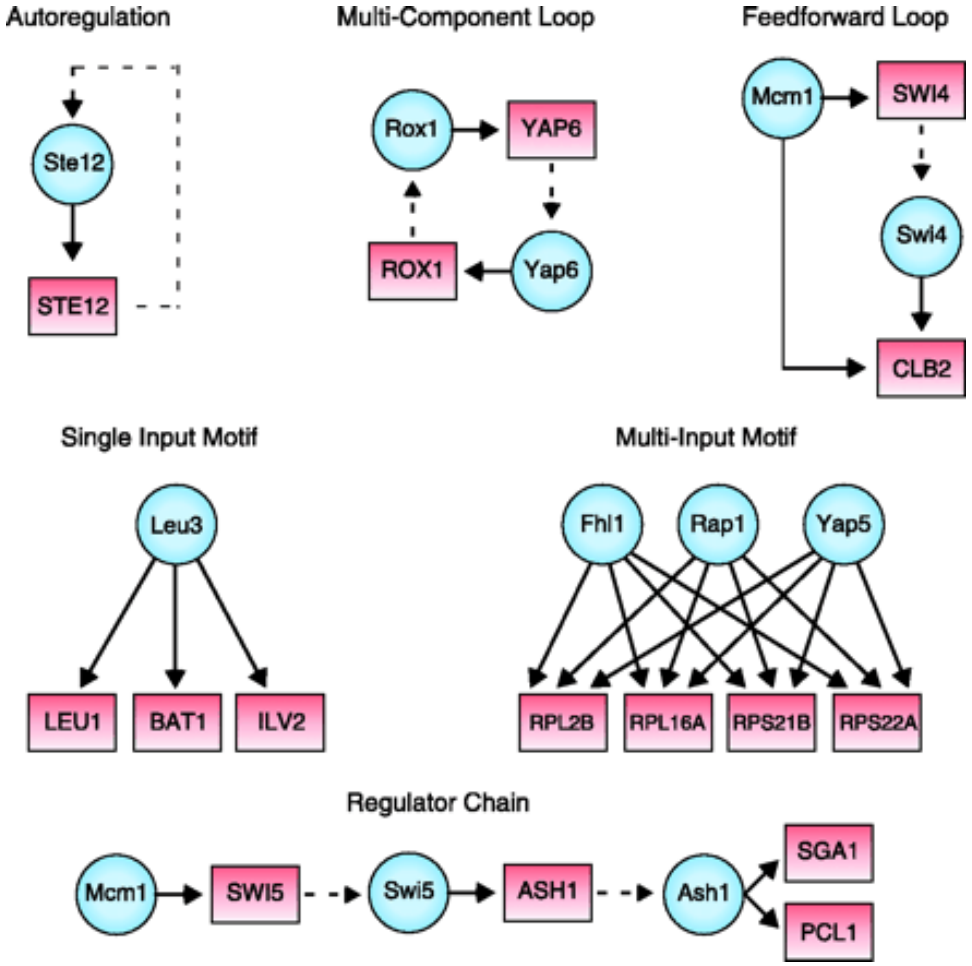
Q: How does the non-cumulative distribution look like in the region where the cumulative distribution is almost horizontal?



R. Albert, I. Albert, G. L. Nakarado, Phys. Rev. E 69, 025103(R) (2004)

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Electronic circuits (forward logic chips)					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

Transcriptional regulatory motifs



Regulators (TFs), blue circles

Genes, red rectangles

Dashed edges mean translation

Feedforward loop: convergent direct and indirect regulation; noise filter

Single input motif: one TF regulates several genes; temporal program

Multi-input motif: combinatorial regulation

Lee et al, Science 298, 799 (2002)

Mixing patterns in networks

Mixing in social networks

- assortative: people prefer to associate with others who are like them
- disassortative: people prefer to associate with others who are different

Mixing with respect of node degree:

- assortative: high degree nodes tend to be connected to high degree nodes
- disassortative: high degree nodes tend to be connected to low degree nodes

Focus on edge i, denote the excess in-degree of its starting point with j_i and the excess out-degree of its endpoint with k_i

Mixing is quantified by the correlation between j_i and k_i over all i

$$r = \frac{\sum_i j_i k_i - \sum_i j_i \sum_{i'} k_{i'} \Big/ N}{\left(\sum_i j_i^2 - (\sum_i j_i)^2 \Big/ N\right)^{0.5} \left(\sum_i k_i^2 - (\sum_i k_i)^2 \Big/ N\right)^{0.5}}$$

Positive correlation - assortative, Negative correlation - disassortative

	network	type	size n	assortativity r	error σ_r	ref.
social	physics coauthorship	undirected	52909	0.363	0.002	a
	biology coauthorship	undirected	1520251	0.127	0.0004	a
	mathematics coauthorship	undirected	253339	0.120	0.002	b
	film actor collaborations	undirected	449913	0.208	0.0002	c
	company directors	undirected	7673	0.276	0.004	d
	student relationships	undirected	573	−0.029	0.037	e
	email address books	directed	16881	0.092	0.004	f
technological	power grid	undirected	4941	−0.003	0.013	g
	Internet	undirected	10697	−0.189	0.002	h
	World-Wide Web	directed	269504	−0.067	0.0002	i
	software dependencies	directed	3162	−0.016	0.020	j
biological	protein interactions	undirected	2115	−0.156	0.010	k
	metabolic network	undirected	765	−0.240	0.007	l
	neural network	directed	307	−0.226	0.016	m
	marine food web	directed	134	−0.263	0.037	n
	freshwater food web	directed	92	−0.326	0.031	o

Social networks tend to be assortative, technological and biological networks tend to be disassortative.

Possible causes of assortativity: group affiliation, attraction of similars;
Possible causes of disassortativity: service relationships (e.g. directories), representation as simple graphs.

M. E. J. Newman, Phys. Rev. E (2003)

average distance of reachable pairs

<k>

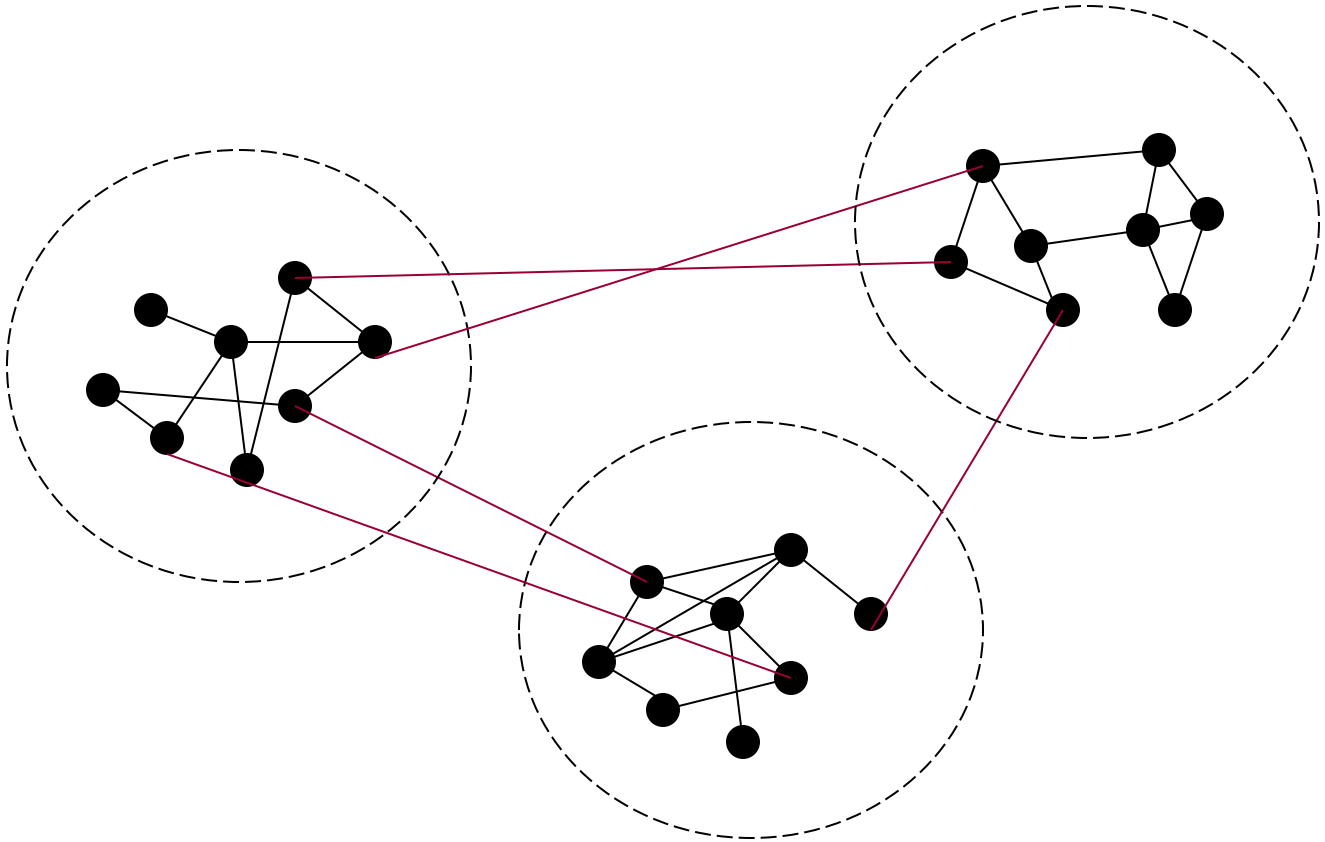
	Network	Type	n	m	c	S	ℓ	α	C	C_{WS}	r	Ref(s).
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20	0.78	0.208	16,323
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59	0.88	0.276	88,253
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15	0.34	0.120	89,146
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45	0.56	0.363	234,236
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088	0.60	0.127	234,236
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1				9,10
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0		0.16		103
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17	0.13	0.092	248
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	−0.029	34
	Sexual contacts	Undirected	2 810					3.2				197,198
Information	WWW nd.edu	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11	0.29	−0.067	13,28
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7				56
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–				280
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13	0.15	0.157	184
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7		0.44		97,116
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035	0.39	−0.189	66,111
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10	0.080	−0.003	323
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–		0.69	−0.033	294
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070	0.082	−0.016	239
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033	0.012	−0.119	315
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010	0.030	−0.154	115
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012	0.011	−0.366	6,282
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	−0.240	166
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	−0.156	164
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	−0.263	160
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	−0.326	209
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	−0.226	323,328

M Newman, Networks (2010)

fraction of nodes in largest (weakly) connected component

Community structure in networks

- Many real-world networks exhibit community structure (also called modularity).
- Intuitively modularity can be defined as the existence of subgraphs that are densely intra-connected but sparsely inter-connected.



Definitions of a community

- Cliques (completely connected subgraphs)
- Chain of cliques – adjacent cliques share every node except one

- k-clan – diameter (largest path length) is $\leq k$

- Definitions using the edges inside and outside a presumed community

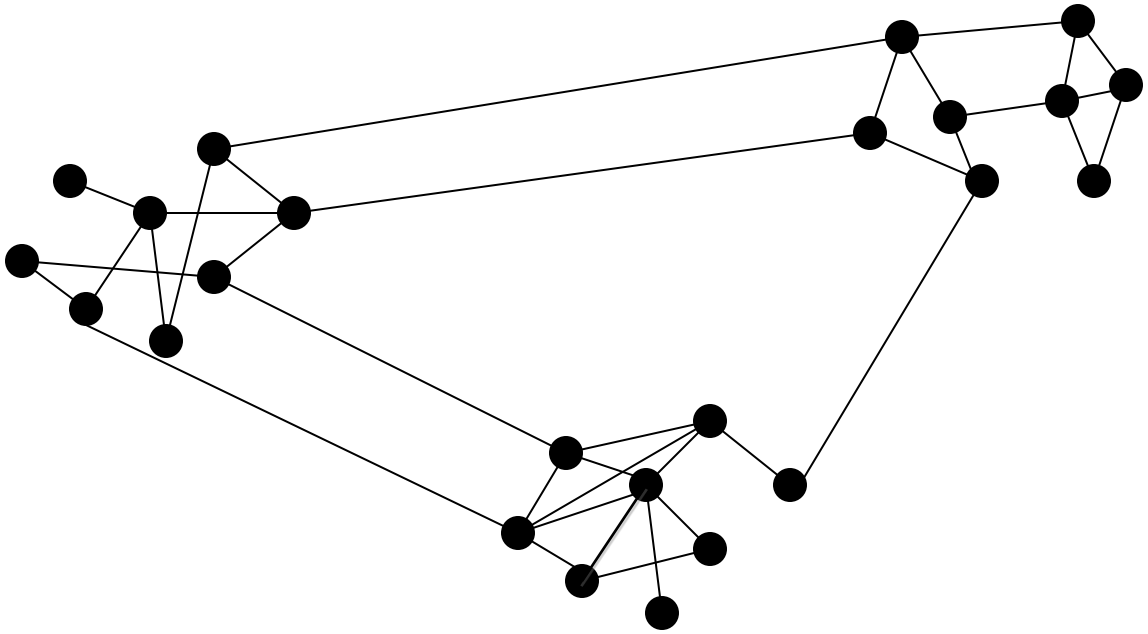
  k_i^{in} – edges of node i that stay inside the community

  k_i^{out} – edges of node i that go outside of the community

  - Strong community: $k_i^{in} \geq k_i^{out}$ for every node i in the community
  - Weak community: $\sum_i k_i^{in} \geq \sum_i k_i^{out}$, where the sum is over nodes in the community

F. Radicchi et al., PNAS 101, 2658 (2004).

- Find cliques, chains of cliques, 2- and 3-clans, strong and weak communities in the graph

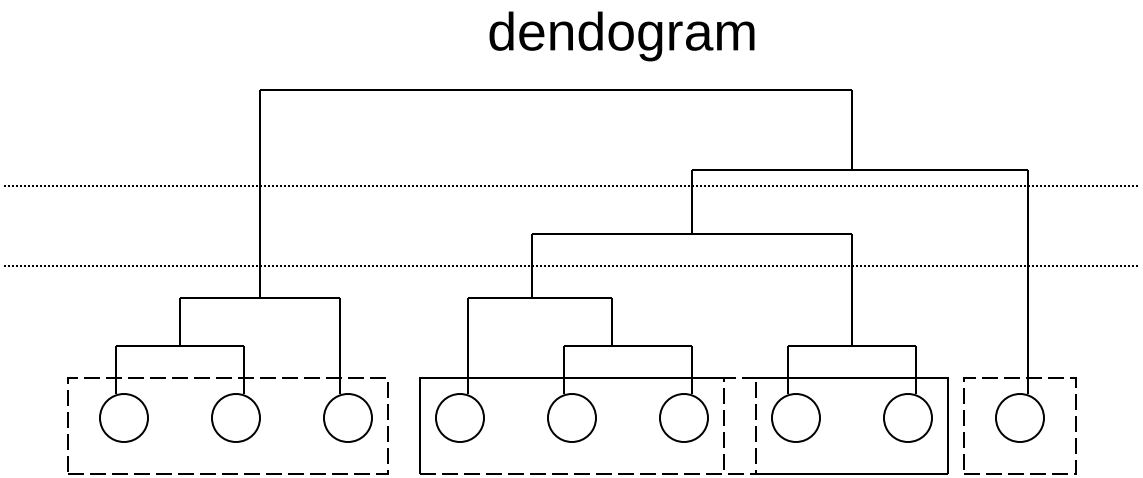


Community Detecting Algorithms

- Most (but not all) methods assume non-overlapping communities
- Two main families of methods:
  - Agglomerative (bottom up)
  - Divisive (top down)
- Several implemented methods
  - Cytoscape has several plugins such as MCODE
  - CFinder

Agglomerative method: hierarchical clustering

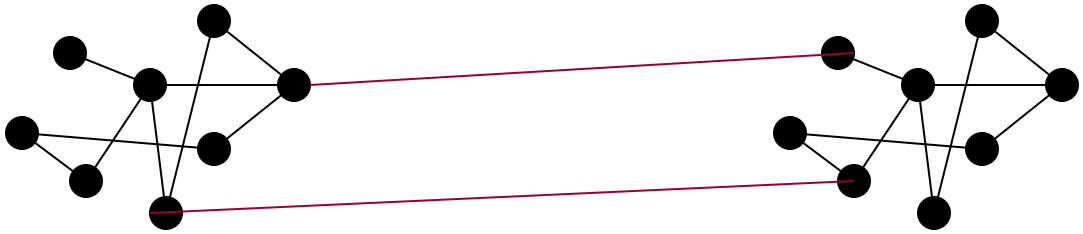
- Calculate a weight (connectivity measure) W_{ij} for every pair *i*, *j* of vertices
  - Example of weight: number of node-independent paths between *i* and *j*.
- Start with each node as a separate community
- Unite the highest-weight node pair(s)
- Calculate the weights between the newly formed community(ies) as averages over the nodes in the community
- Repeat



Divisive method: betweenness centrality algorithm

- Betweenness centrality of an edge is the number of shortest paths between pairs of vertices that run along it
- Algorithm:
  - Calculate the betweenness for all edges in the network
  - Remove the edge with highest betweenness
  - Recalculate the betweenness for all edges affected by the removal
  - Repeat

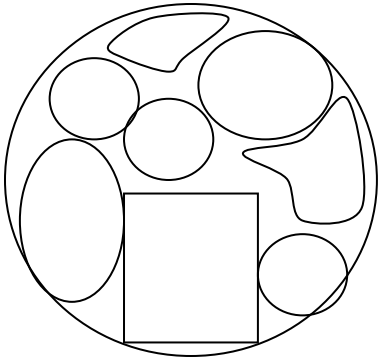
M.E.J. Newman, Phys. Rev. E 69, 066133, 2004

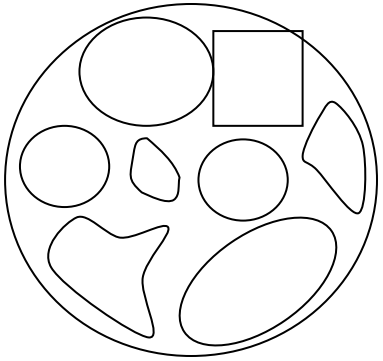


This algorithm also leads to a dendogram

Q: When is it most meaningful to stop?

Strength of communities





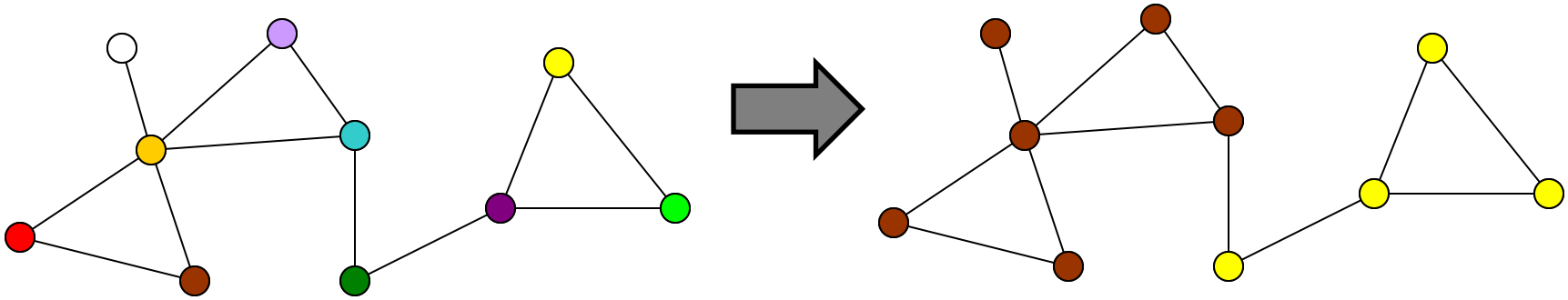
To check if a particular division is meaningful, we can determine the modularity measure Q, defined as the fraction of edges that fall within communities, minus the expected value of the same quantity if edges fall at random without regard for the community structure.

If Q=0, implies the division gives no more within-community edges than would be expected by random chance.
Q>0 indicates a significant community structure.
The higher Q, the better the proposed community structure.

M. Girvan and M.E.J. Newman, PNAS 99 (2002).

Label propagation

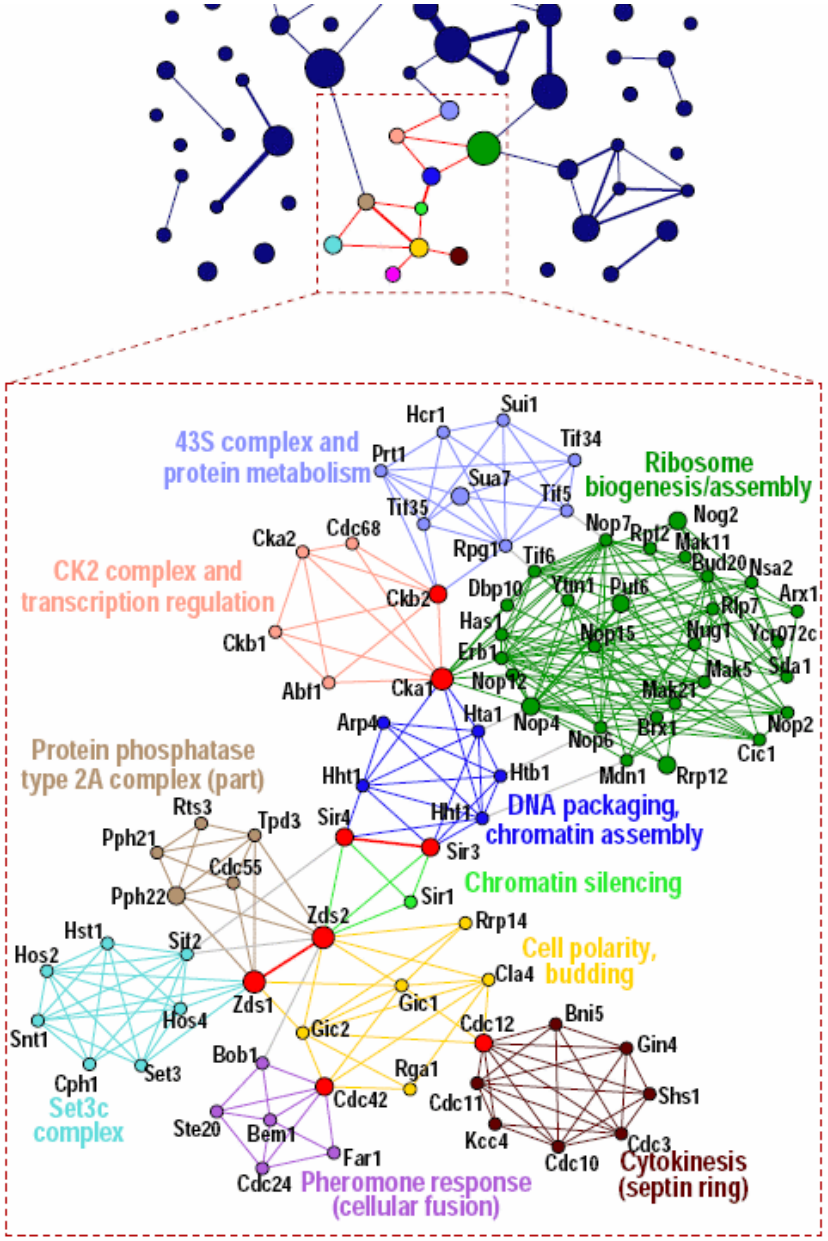
- Each node is initialized with a separate label (is its own community)
- Node labels updated in asynchronous rounds
- Label adoption condition: join the community to which the most adjacent nodes belong. Ties are broken randomly.
- Stop when each node is in the community where most of its neighbors are.
- No unique solution, but solutions are similar to each other.
- Faster and as efficient as other algorithms



U.N. Raghavan, R. Albert, S. Kumara PRE 76, 036106 (2007).

Clique percolation

- Idea: a community can be interpreted as a union of cliques that share nodes
- **k-clique-community** is the union of all k-cliques that can be reached from each other through a series of adjacent k-cliques.
- Two k-cliques are adjacent if they share k-1 nodes.
- k-clique-communities can form meta-nodes in a higher level network.



Palla *et. al.*, Nature 435, 814-818 (2005)